Genomic resources for Arabidopsis research at NCBI

National Center for Biotechnology Information . National Library of Medicine . National Institutes of Health

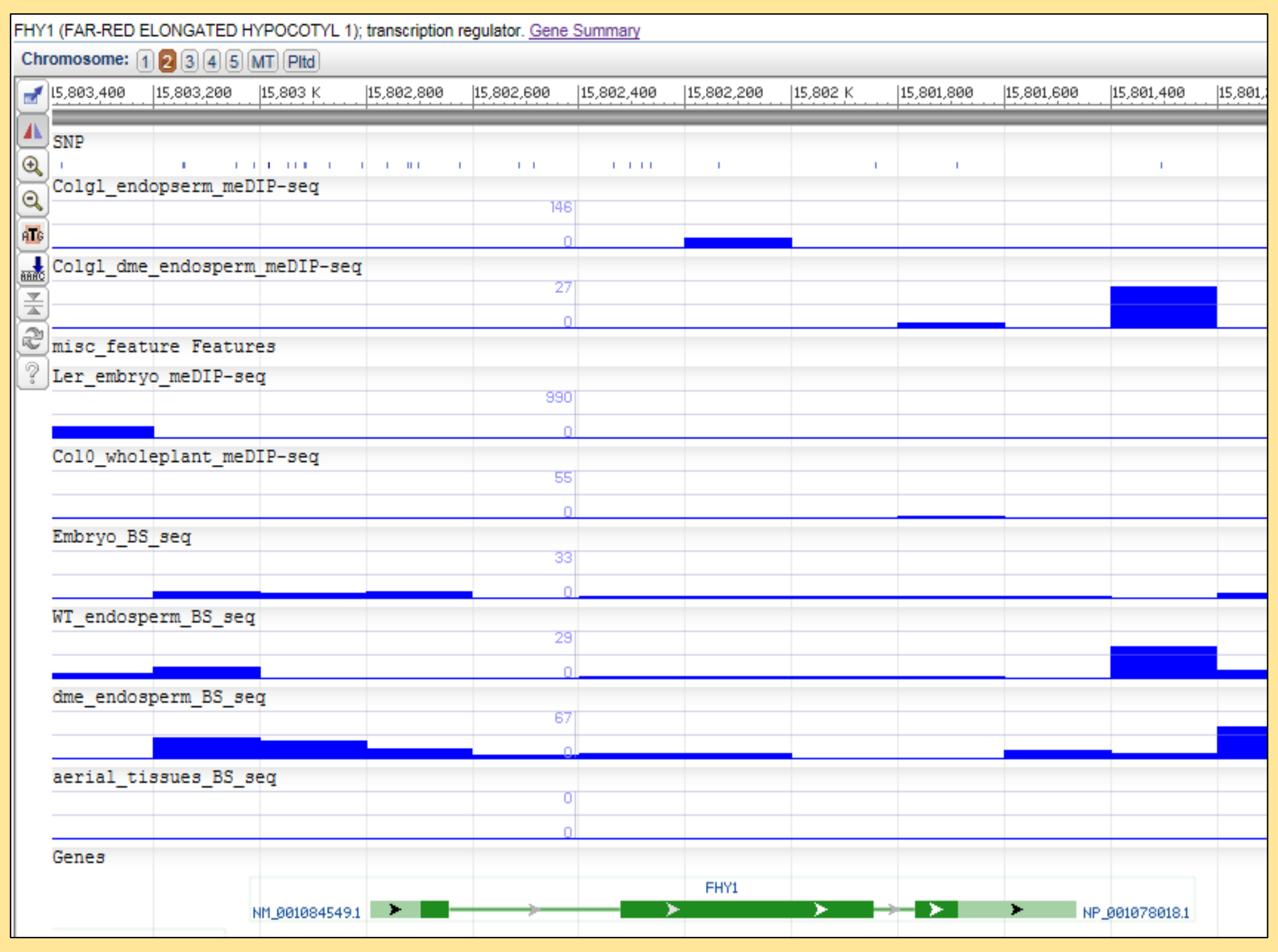
Anjana R. Vatsan, Terence D. Murphy, Valerie Schneider, Brian Smith-White, Tatiana Tatusova, Kim D. Pruitt

info@ncbi.nlm.nih.gov

ABSTRACT

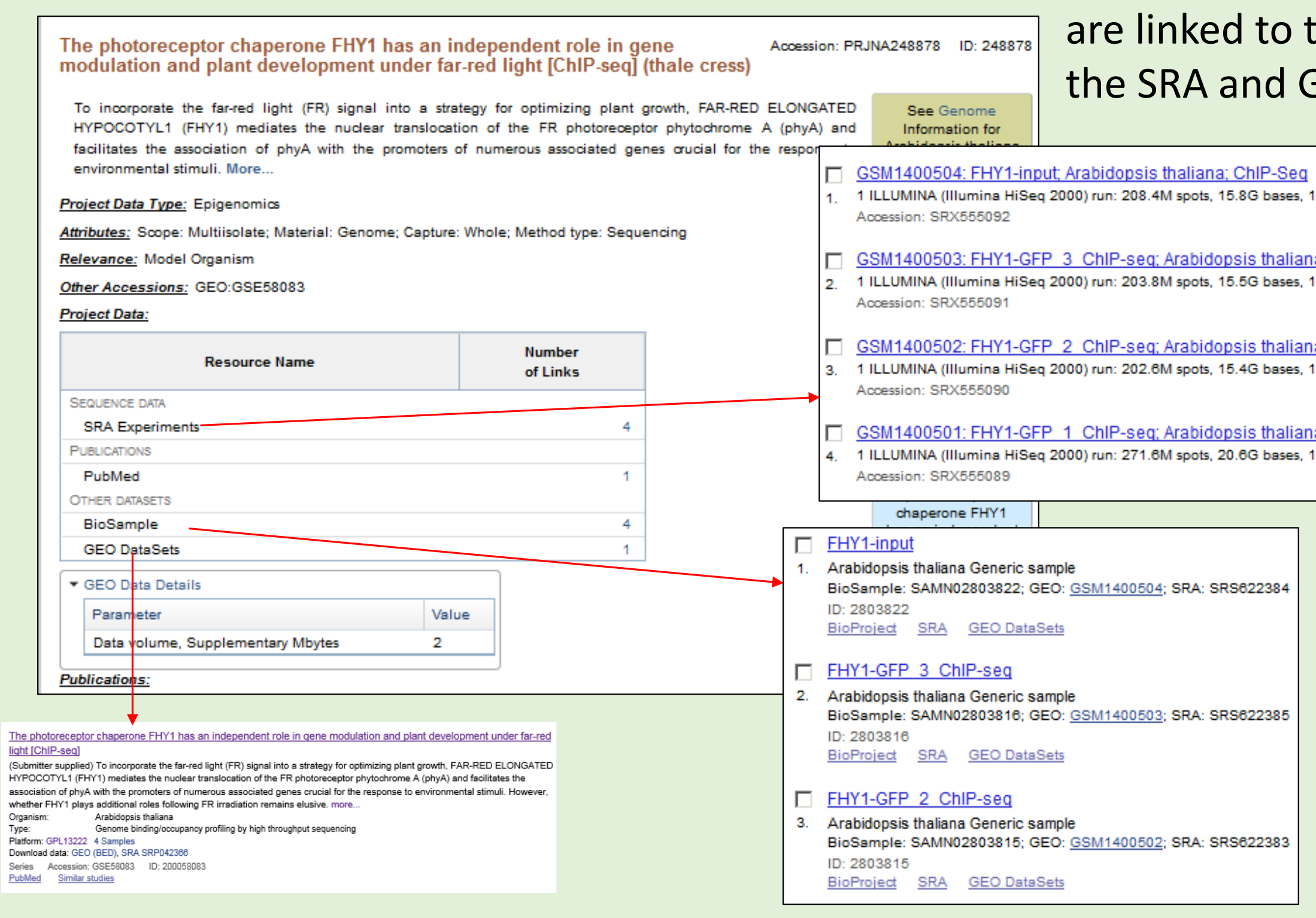
NCBI's Entrez system is an integrated search and retrieval system that provides access to a diverse set of 39 molecular and literature databases. Arabidopsis thaliana data is available in several NCBI resources ranging from primary data submissions, meta-data resources, and core resources that provide organized views of genomes, genes, maps, and sequences. Data gets submitted to one of NCBI's archival databases - GenBank, Assembly, BioProject, BioSample, GEO, SRA, TSA, dbSNP, PopSet, Epigenomics - and then through a combination of manual curation and computational methods is integrated into NCBI's secondary databases like, BLAST, CloneDB, Gene, RefSeq, Map Viewer, Genome, HomoloGene, UniGene, and Protein Clusters. Over 2100 Arabidopsis BioProjects are registered with NCBI; these projects reflect the areas of ongoing research ranging from large scale sequencing projects to expression studies related to tissue type or environmental conditions or epigenomics. Data is then made available through several Entrez databases. NCBI's Entrez search system allows for searches between databases including the PubMed database. The presentation will include a summary of available data of different types and will include information on data access.

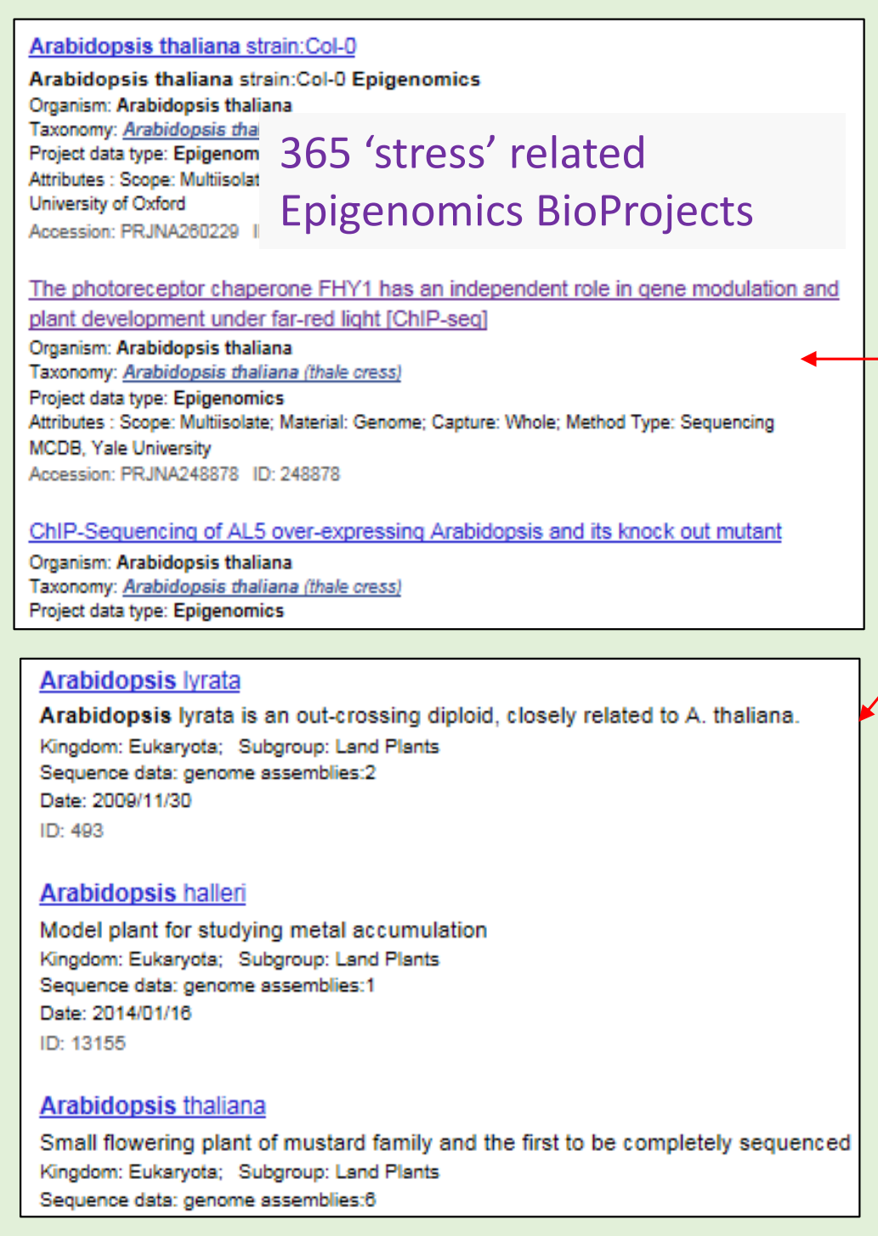
A subset of **epigenetics**-specific data from the GEO and SRA databases is processed and mapped to genomic coordinates to generate tracks that provide a visual representation of the data.



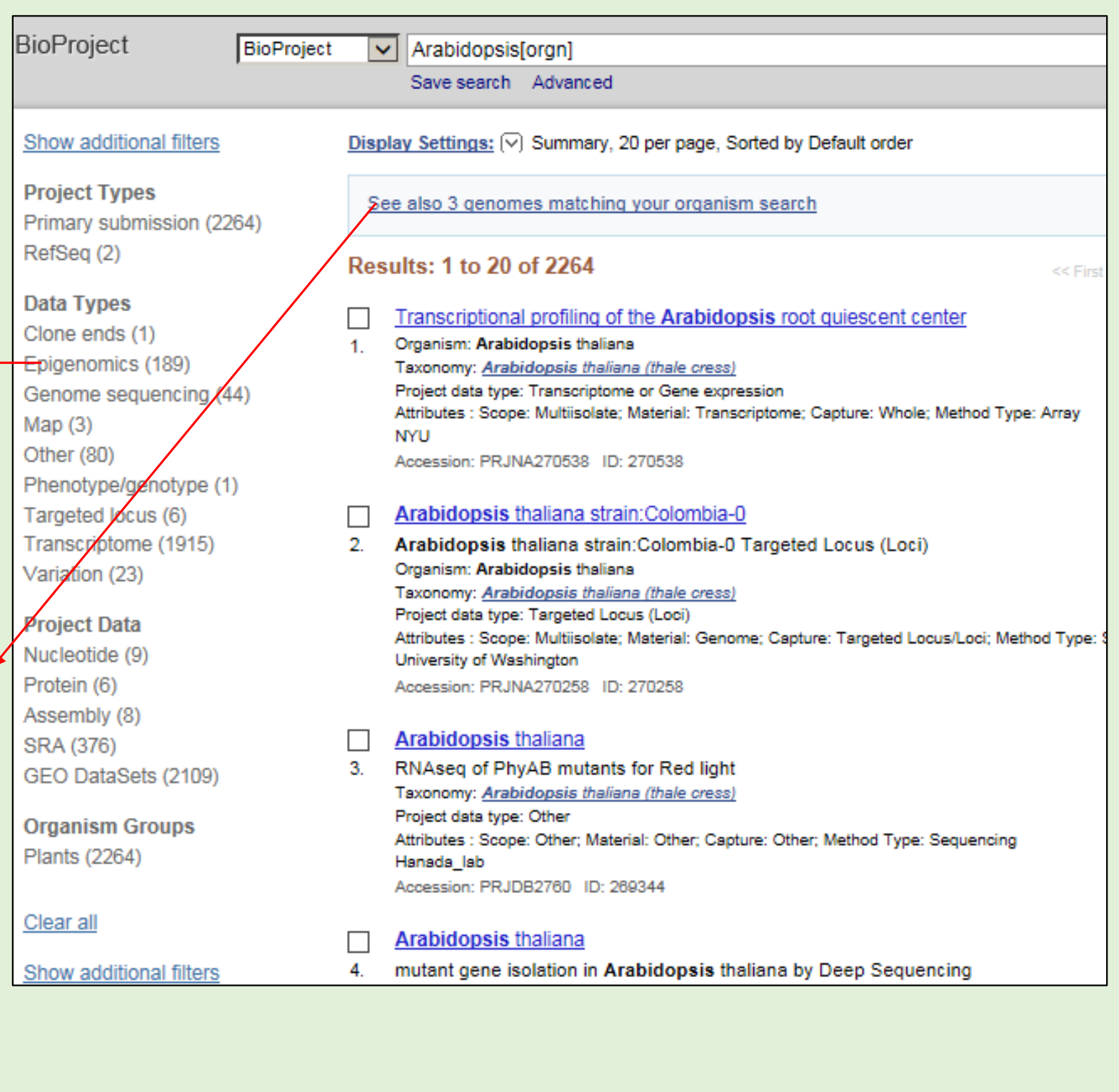
The **BioProjects** database describes large-scale research efforts, ranging from genome and transcriptome sequencing projects to epigenomic and variation analyses. It provides an organizational framework to access data across multiple resources and multiple submission time points. As of now, close to 2,300 BioProjects are registered for Arabidopsis of which 2,234 are for A. thaliana.

The **Epigenomics** database is a comprehensive resource for whole-genome epigenetic data sets. Epigenomics BioProjects are linked to the raw data in the SRA and GEO databases.

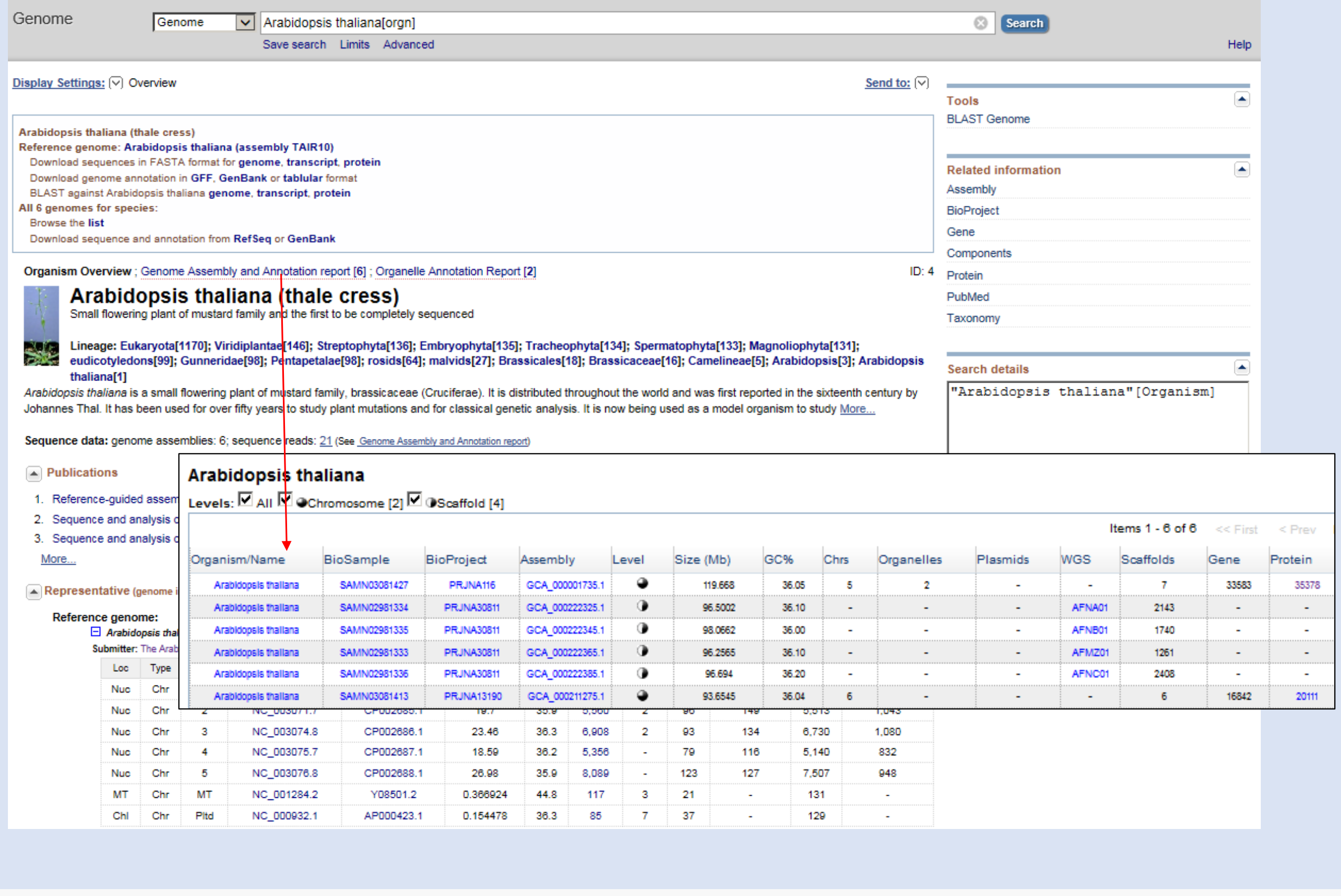




365 'stress' related Epigenomics BioProjects

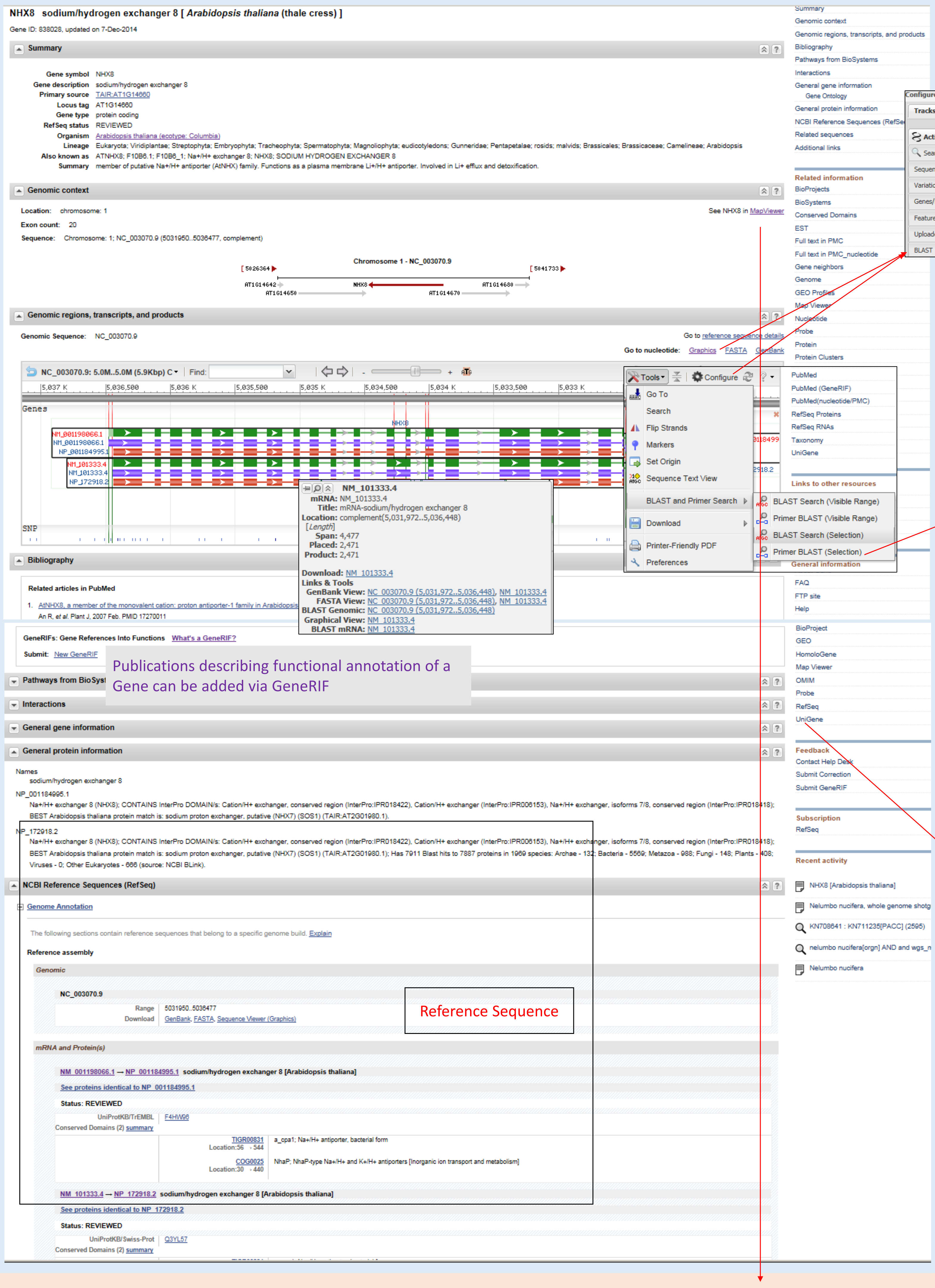


The **Genome** database provides organism specific genomic information with links to assembly and annotation information.



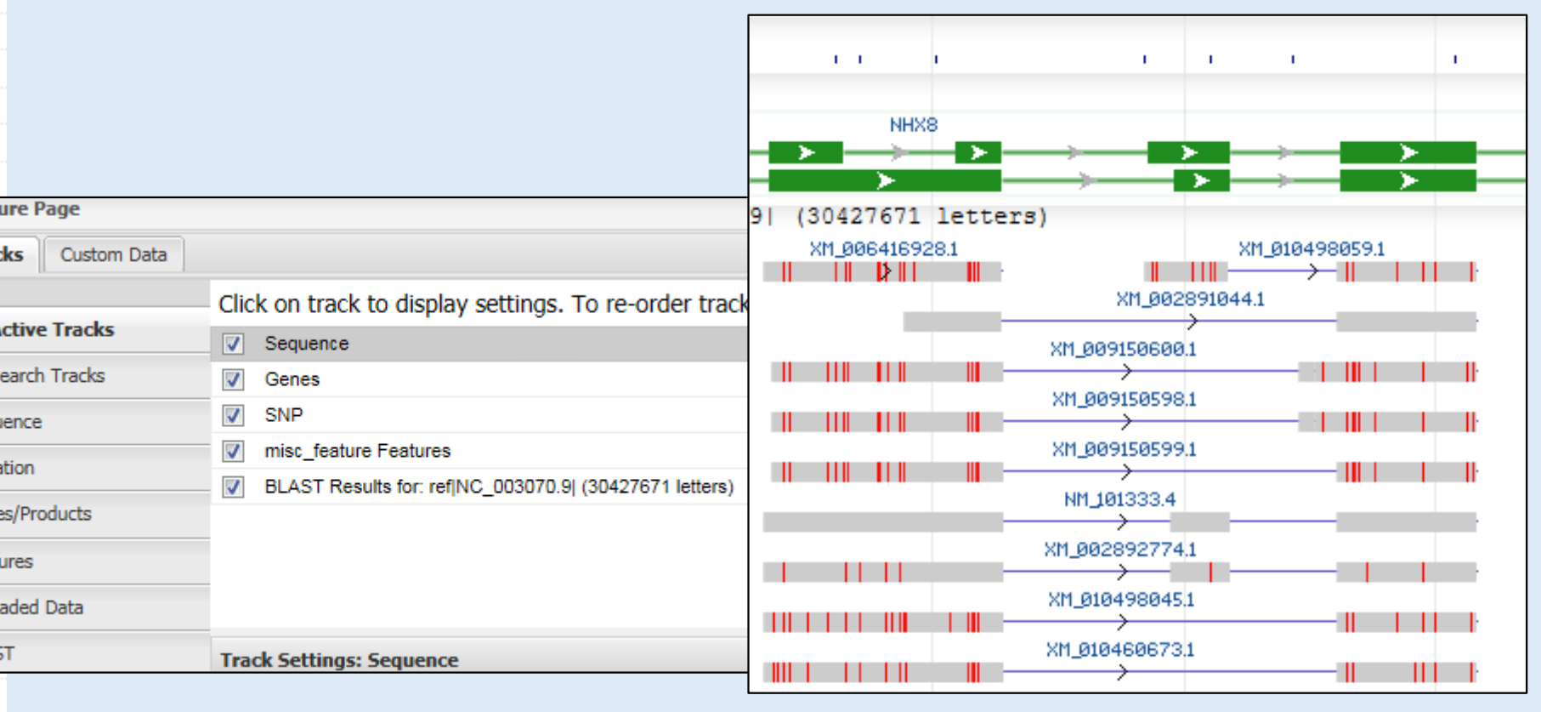
The **Gene** database serves as a central hub for gene-centric information. It allows for rapid access to Reference Sequences and graphical displays of genomic maps.

The RefSeqs for a Gene record are presented in an interactive and customizable graphical **Sequence Viewer**. Regions of interest can be explored using the "Tools" menu. Sequence can be viewed, downloaded or BLAST aligned. Configure button allows for various tracks to be aligned and viewed in a single window.

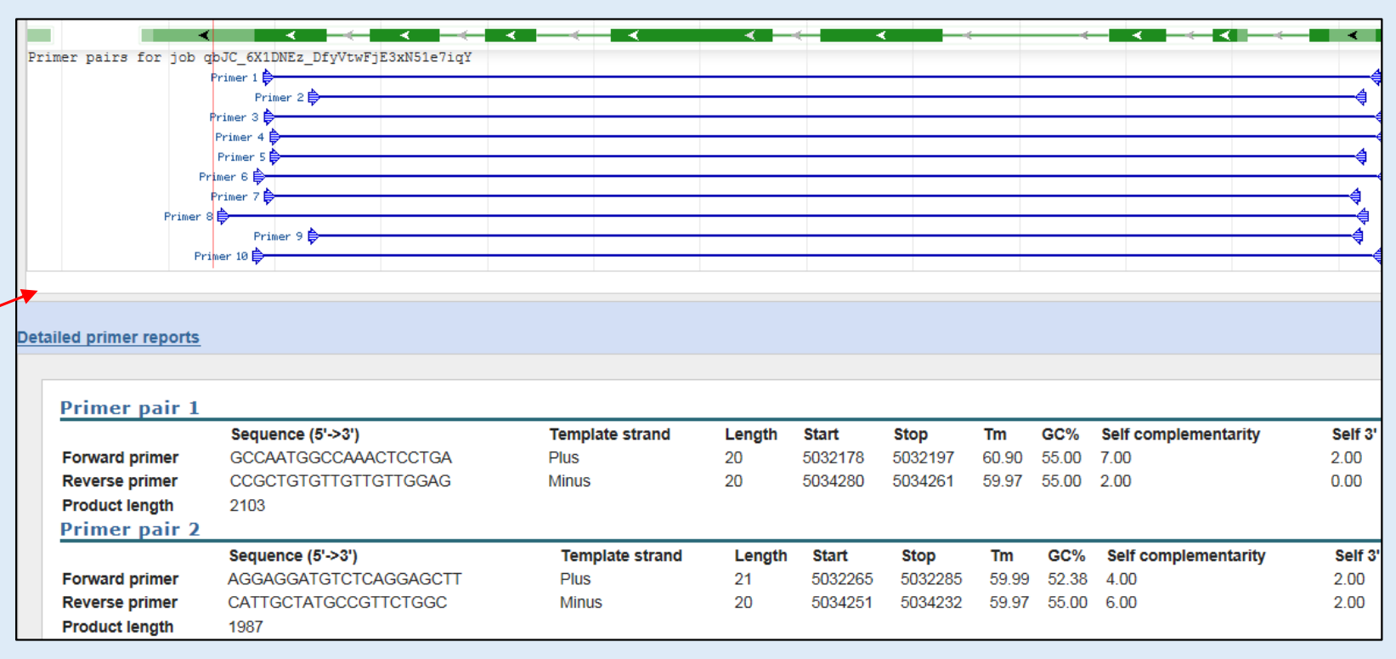


Publications describing functional annotation of a Gene can be added via GeneRIF

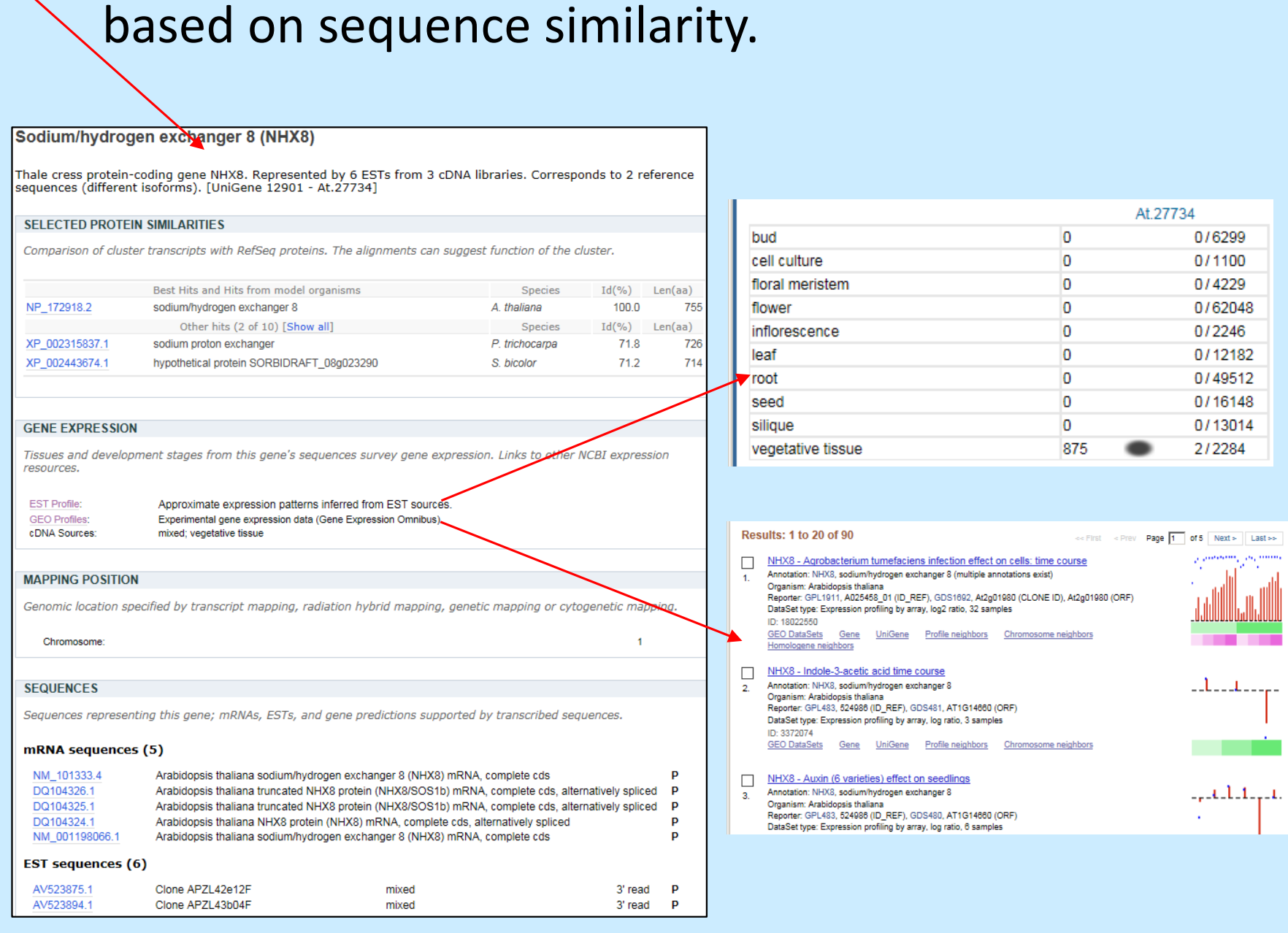
Reference Sequence



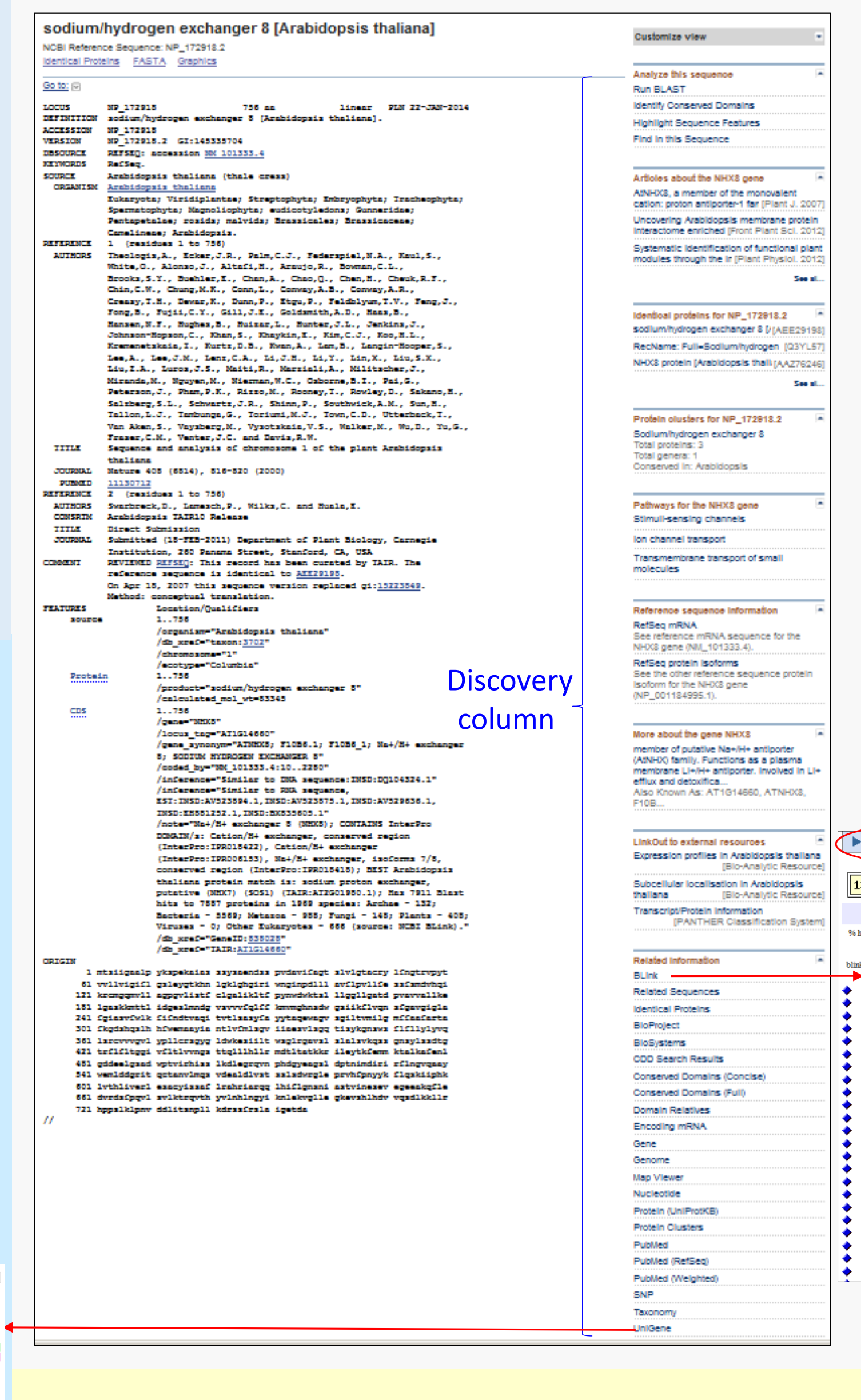
Primer BLAST allows for the selection of primer sequences within a selected region.



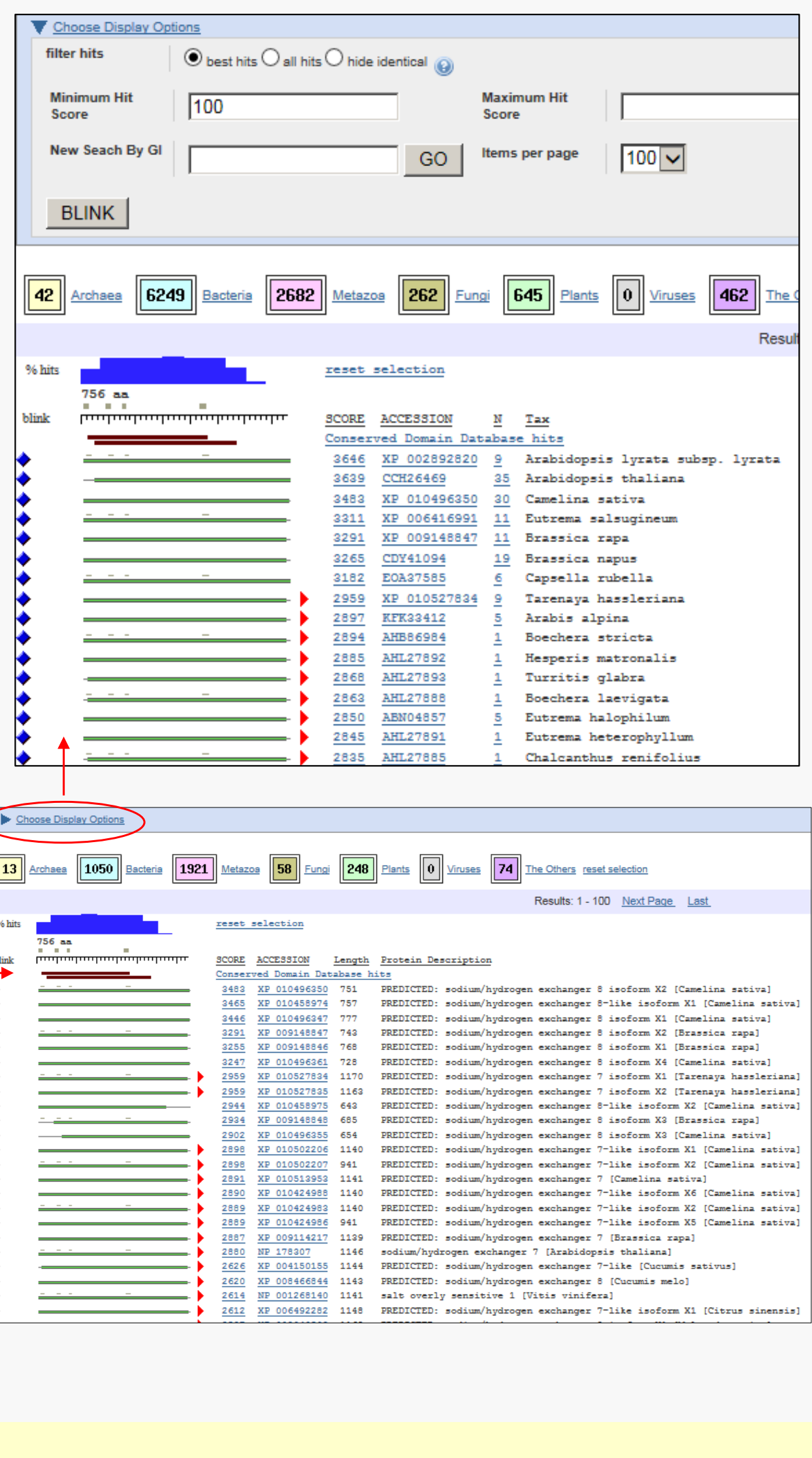
RefSeq records are linked to the **UniGene** database that connects expressed sequences from GenBank and RefSeqs into gene clusters based on sequence similarity.



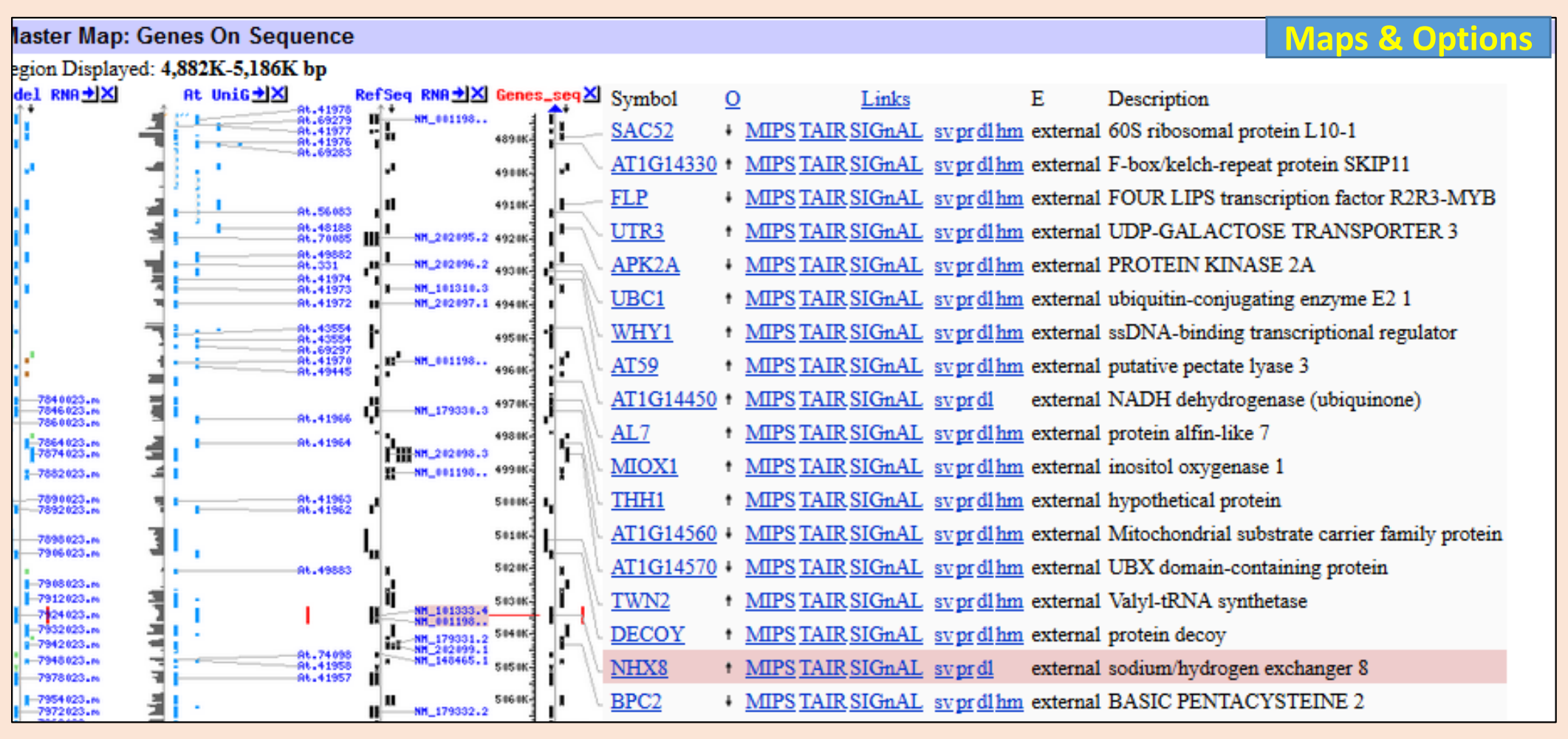
Reference Sequence (RefSeq) protein records are in the **Protein** database. The Discovery column on each RefSeq record provides access to a wealth of information from related databases. Homologs can be identified following the "Blink" in the protein record. "Choose Display Options" provides filters to show single or unique protein from each organism.



Discovery column

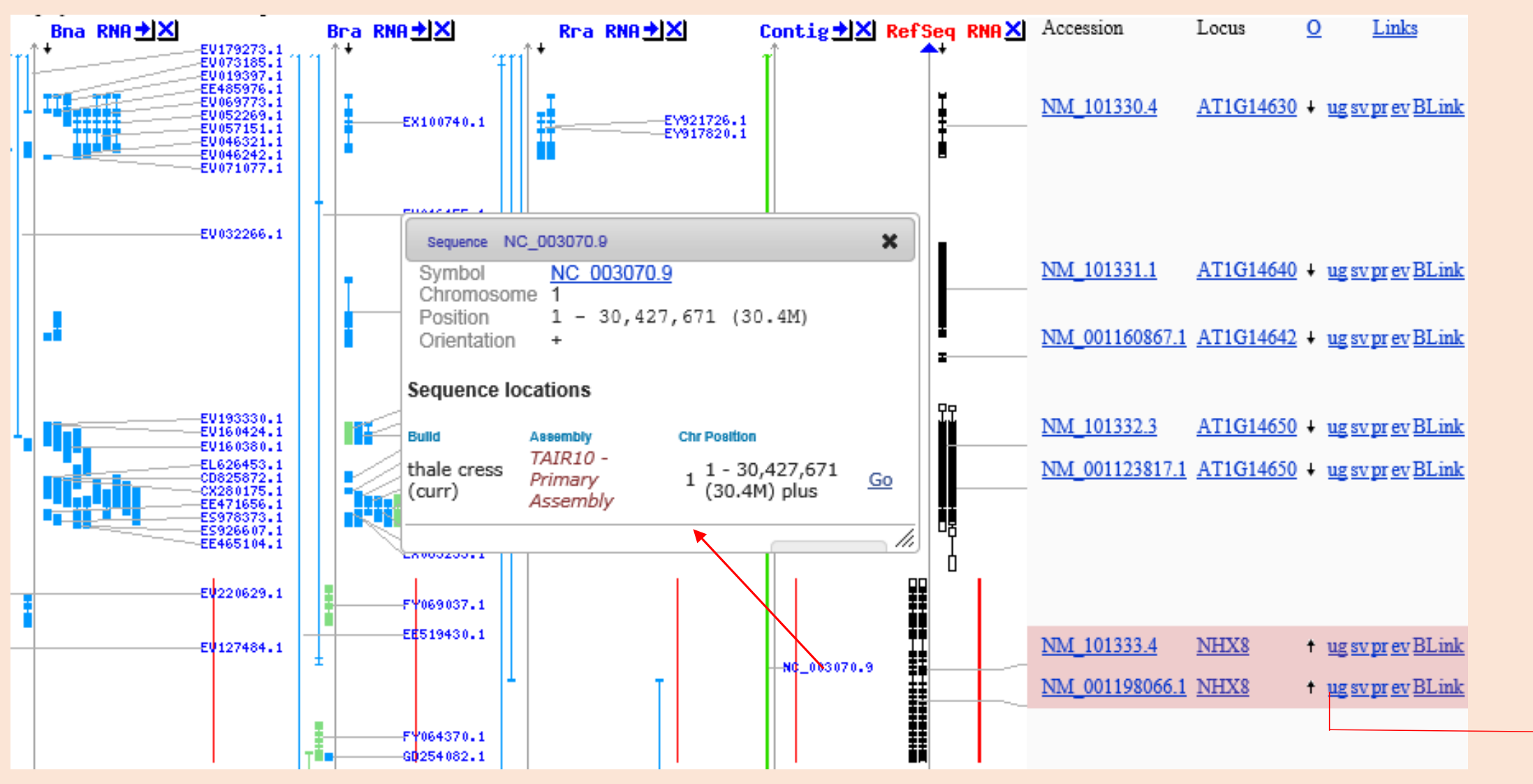


The **Map Viewer** is the primary graphical display tool for eukaryotic genomes. Both sequence and non-sequence maps can be displayed, aligned and searched.

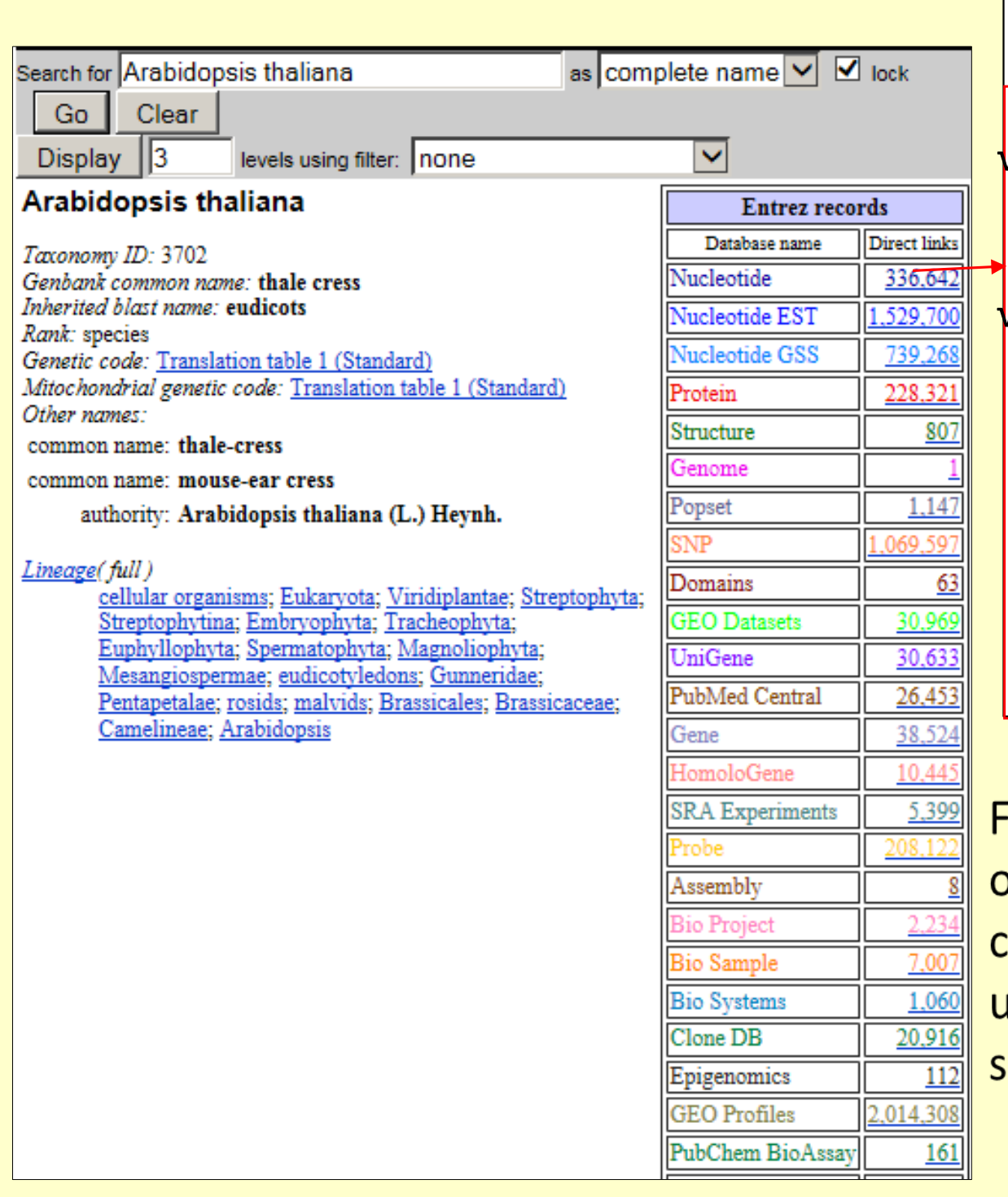


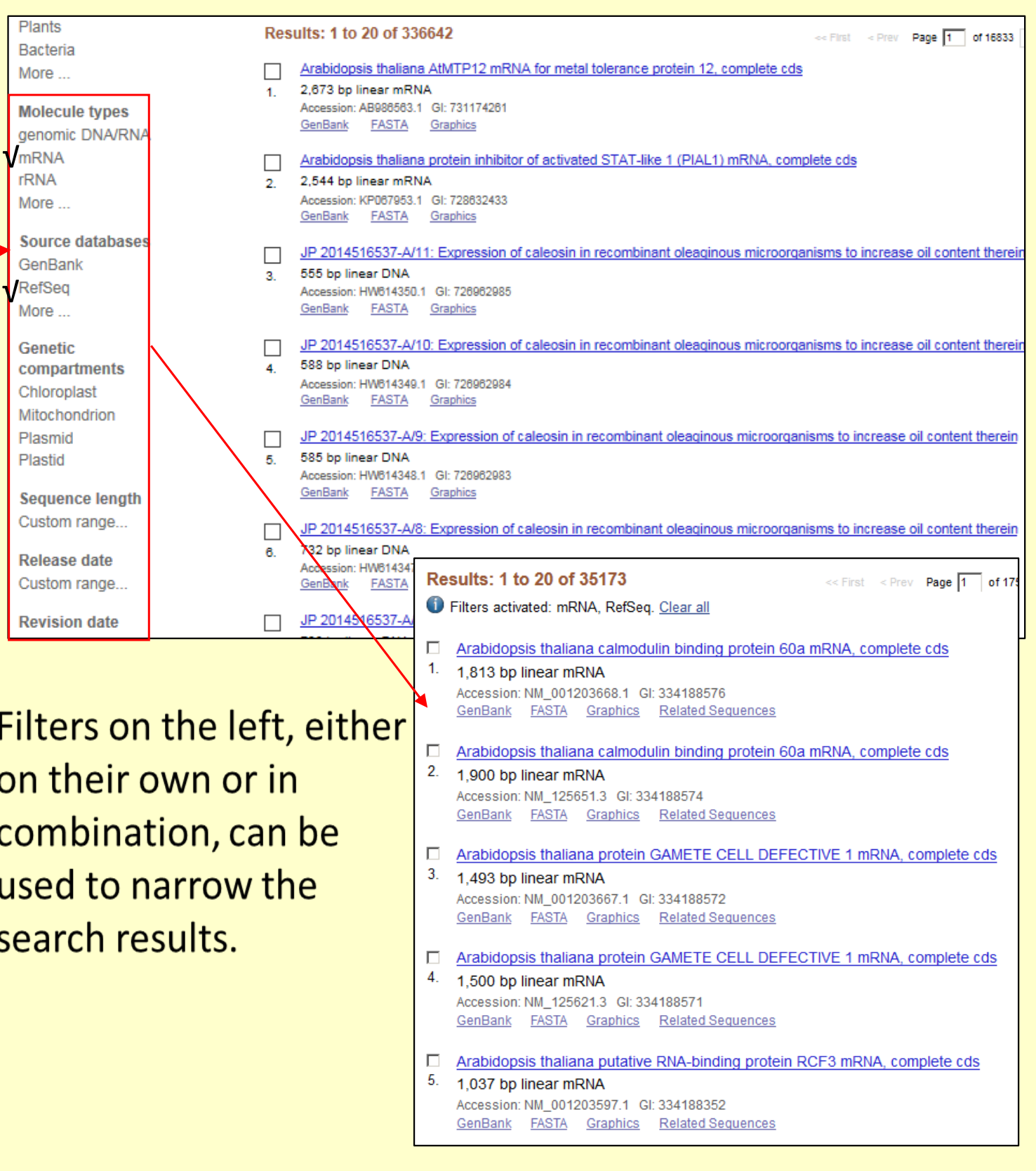
Maps & Options

The **"Maps and Options"** dialog allows for the display of multiple maps. Genome assembly information can be found by selecting the Contig map. The image below displays the A. thaliana, B. napus, B. rapa and R. raphanistrum transcript maps aligned to A. thaliana genome .



The **Taxonomy** database not only provides phylogenetic lineages of more than 160,000 organisms that have data in the various NCBI databases, but also provides a quick overview and access to the data for a particular organism. The nucleotide sequences for A. thaliana are present in three of the databases: Nucleotide, EST, and GSS. Reference Sequence RNA and genomic records are in the Nucleotide database.





Filters on the left, either on their own or in combination, can be used to narrow the search results.